Exact formulas for the variance of several balance indices under the Yule model

Gabriel Cardona, Arnau Mir, and Francesc Rosselló

Dept. of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma.
{gabriel.cardona,arnau.mir,cesc.rossello}@uib.es

Abstract. One of the main applications of balance indices is in tests of null models of evolutionary processes. The knowledge of an exact formula for a statistic of a balance index, holding for any number n of leaves, is necessary in order to use this statistic in tests of this kind involving trees of any size. In this paper we obtain exact formulas for the variance under the Yule model of the Sackin index and the total cophenetic index of binary rooted phylogenetic trees with n leaves. We also obtain the covariance of these indices.

1 Introduction

One of the most thoroughly studied properties of the topology of phylogenetic trees is their symmetry, that is, the degree to which both children of each internal node tend to have the same number of descendant taxa. The symmetry of a tree is usually quantified by means of *balance indices*. Many such indices have been proposed so far in the literature [5, Chap. 33]. One of the most popular is *Sackin's index* S [15], which is defined as the sum of the depths of all leaves in the tree. We have recently proposed an extension of Sackin's index, the *total cophenetic index* Φ [11]: the sum, over all pairs of different leaves of the tree, of the depth of their least common ancestor. The main advantages of Φ over S are that it has a larger range of values and a smaller probability of ties. Moreover, Φ retains other good properties of S: it makes sense for not necessarily fully resolved phylogenetic trees (unlike another popular balance index, Colless' [4]), it can be computed in linear time, and the statistical properties of its distribution of values can be studied under different stochastic models of evolution, like for instance the Yule [7,21] and the uniform [3,14,18] models. This last property is relevant because one of the main applications of balance indices is their use as tools to test stochastic models of evolution [12,16].

Exact formulas for the expected values of S and Φ on the space $\mathcal{T}_n$ of fully resolved rooted phylogenetic trees with n leaves have been published for the Yule and the uniform models. In particular, if we denote by H_n the n-th *harmonic number*, i.e.,

$$H_n = \sum_{i=1}^{n} \frac{1}{i},$$

these expected values under the Yule model on $\mathcal{T}_n$ are, respectively,

$$\begin{array}{ll} E_Y(S_n) = 2n(H_n - 1) & [8] \\ E_Y(\Phi_n) = n(n-1) - 2n(H_n - 1) & [11] \end{array}$$

As we already pointed out [11], these formulas imply that the expected value under the Yule model of the sum $\overline{\Phi} = S + \Phi$ on $\mathcal{T}_n$ is $n(n-1)$, a quite simpler expression than those for $E_Y(S_n)$ or $E_Y(\Phi_n)$. This index $\overline{\Phi}$ has the same good properties of Φ, but the formulas for its statistics under the Yule model tend to be simpler than the corresponding formulas for other indices. We shall find here another example of this fact: the variance.

The goal of this paper is to provide exact formulas for the variance of S, Φ and $\overline{\Phi}$, and the covariances between S and the other two, on $\mathcal{T}_n$ and under the Yule model. The variance of S on $\mathcal{T}_n$ under this model was known so far only for its limit distribution when $n \to \infty$ [1], being

$$\sigma_Y^2(S_n) \sim \Big(7 - \frac{2\pi^2}{3}\Big) n^2.$$

Also, Rogers [13] found a recursive formula for the moment-generating functions of S under this model, which allowed him to compute $\sigma_Y^2(S_n)$ for $n = 1, \ldots, 50$, but he did not obtain any explicit exact formula for this variance.

In this paper we obtain the following exact formulas for these variances:

$$\begin{aligned}
\sigma_Y^2(S_n) &= 7n^2 - 4n^2 H_n^{(2)} - 2nH_n - n\\
\sigma_Y^2(\Phi_n) &= \frac{1}{12}(n^4 - 10n^3 + 131n^2 - 2n) - 6nH_n - 4nH_n^2 - 4n(n-1)H_n^{(2)}\\
\sigma_Y^2(\overline{\Phi}_n) &= 2\binom{n}{4}
\end{aligned}$$

where $H_n^{(2)} = \sum\limits_{i=1}^{n} 1/i^2$. We also obtain the following exact formulas for the covariances, under the Yule model, of S_n and $\overline{\Phi}_n, \Phi_n$:

$$\begin{aligned}
cov_Y(S_n, \Phi_n) &= 4n(H_n^{(2)} + H_n) + \frac{1}{6} n(n^2 - 51n + 2)\\
cov_Y(S_n, \overline{\Phi}_n) &= 2nH_n + \frac{1}{6} n(n^2 - 9n - 4)
\end{aligned}$$

These formulas are valid for any number n of leaves, and therefore they can be used in a meaningful way in tests involving trees of any size. The proofs of all these formulas consist of elementary, although technically involved, algebraic computations.

The rest of this paper is organized as follows. In a first section on Preliminaries we gather some notations and conventions on phylogenetic trees and some lemmas on probabilities of trees under the Yule model and on harmonic numbers. In the next section, we establish a recursive formula for the expected value under the Yule model of the square of a balance index satisfying a certain kind or recursion (a *recursive shape index* [10]) that lies at the basis of all our computations. Then, we devote a series of sections to compute the variances of S, Φ, $\overline{\Phi}$ and the covariance of S with Φ and $\overline{\Phi}$, respectively. These sections consist of long and tedious algebraic computations, without any interest beyond the fact that they prove the formulas announced above. We close the paper with a section on Conclusions and Discussion.

2 Preliminaries

2.1 Phylogenetic trees

In this paper, by a *phylogenetic tree* on a set S of taxa we mean a binary rooted tree with its leaves bijectively labeled in the set S. We shall always understand a phylogenetic tree as a directed graph, with its arcs pointing away from the root. To simplify the language, we shall always identify a leaf of a phylogenetic tree with its label. We shall also use the term *phylogenetic tree with n leaves* to refer to a phylogenetic tree on the set $\{1,\dots,n\}$. We shall denote by $\mathcal{T}(S)$ the set of isomorphism classes of phylogenetic trees on a set S of taxa, and by $\mathcal{T}_n$ the set $\mathcal{T}(\{1,\dots,n\})$ of isomorphism classes of phylogenetic trees with n leaves. We shall denote by $V_{int}(T)$ the set of internal nodes of a phylogenetic tree T.

Whenever there exists a path from u to v in a phylogenetic tree T, we shall say that v is a *descendant* of u and that u is an *ancestor* of v. The *lowest common ancestor* $\mathrm{LCA}_T(u,v)$ of a pair of nodes u,v of a phylogenetic tree T is the unique common ancestor of them that is a descendant of every other common ancestor of them.

The *depth* $\delta_T(v)$ of a node v in T is the length (in number of arcs) of the unique path from the root r of T to v. The *cophenetic value* [17] of a pair of leaves i,j is the depth of their lowest common ancestor:

$$\varphi_T(i,j)=\delta_T(\mathrm{LCA}_T(i,j)).$$

To simplify the notations at some points, we shall also write $\varphi_T(i,i)$ to denote the depth $\delta_T(i)$ of a leaf i.

Given two phylogenetic trees T,T' on disjoint sets of taxa S,S', respectively, their *tree-sum* is the tree $T\widehat{\ }T'$on $S\cup S'$ obtained by connecting the roots of T and T' to a (new) common root. Every tree with n leaves is obtained as $T_k\widehat{\ }T'_{n-k}$, for some $1\leqslant k\leqslant n-1$, some subset $S_k\subseteq\{1,\dots,n\}$ with k elements, some tree T_k on S_k and some tree T'_{n-k} on $S_k^c=\{1,\dots,n\}\setminus S_k$; actually, every tree T with n leaves is obtained in this way *twice.*

The *Yule*, or *Equal-Rate Markov*, model of evolution [7,21] is a stochastic model of phylogenetic trees' growth. It starts with a node, and at every step a leaf is chosen randomly and uniformly and it is splitted into two leaves. Finally, the labels are assigned randomly and uniformly to the leaves once the desired number of leaves is reached. This corresponds to a model of evolution where, at each step, each currently extant species can give rise, with the same probability, to two new species. Under this model of evolution, different trees with the same number of leaves may have different probabilities. More specifically, if T is a phylogenetic tree with n leaves, and for every internal node z we denote by $\kappa_T(z)$ the number of its descendant leaves, then the probability of T under the Yule model is [2,19]

$$P_Y(T)=\frac{2^{n-1}}{n!}\prod_{v\in V_{int}(T)}\frac{1}{\kappa_T(v)-1}.$$

The following easy lemma on the probability of a tree-sum under the Yule model will be used in our computations.

Lemma 1. *Let $\emptyset\neq S_k\subsetneq\{1,\dots,n\}$ with $|S_k|=k$, let $T_k\in\mathcal{T}(S_k)$ and $T'_{n-k}\in\mathcal{T}(\{1,\dots,n\}\setminus S_k)$. Then*

$$P_Y(T_k\widehat{\ }T'_{n-k})=\frac{2}{(n-1)\binom{n}{k}}P_Y(T_k)P_Y(T'_{n-k})$$

Proof. This equality is a direct consequence of the explicit probabilities of $P_Y(T_k)$, $P_Y(T'_{n-k})$ and $P_Y(T_k\widehat{\ }T'_{n-k})$ and the fact that $V_{int}(T_k\widehat{\ }T'_{n-k})$ is the disjoint union of $V_{int}(T_k)$, $V_{int}(T'_{n-k})$ and the root r of $T_k\widehat{\ }T'_{n-k}$. □

2.2 Harmonic numbers

For every $n \geqslant 1$, let

$$H_n = \sum_{i=1}^{n} \frac{1}{i}, \quad H_n^{(2)} = \sum_{i=1}^{n} \frac{1}{i^2}.$$

Let, moreover, $H_0 = H_0^{(2)} = 0$. H_n is called the n-th *harmonic number*, and $H_n^{(2)}$, the *generalized harmonic number of power* 2. It is known (see, for instance, [6, p. 264]) that

$$\begin{aligned} H_n &= \ln(n) + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + O\Big(\frac{1}{n^3}\Big) \\ H_n^{(2)} &= \frac{\pi^2}{6} - \frac{1}{n} + \frac{1}{2n^2} + O\Big(\frac{1}{n^3}\Big) \end{aligned}$$

where γ is Euler's constant.

The following identities will be used in the proofs of our main results.

Lemma 2. *For every $n \geqslant 2$:*

(1) $\sum_{k=1}^{n-1} H_k = n(H_n - 1)$

(2) $\sum_{k=1}^{n-1} kH_k = \frac{1}{4}n(n-1)(2H_n - 1)$

(3) $\sum_{k=1}^{n-1} k^2H_k = \frac{1}{36}n(n-1)((12n-6)H_n - 4n - 1)$

(4) $\sum_{k=1}^{n-1} \frac{H_k}{k+1} = \frac{1}{2}(H_n^2 - H_n^{(2)})$

(5) $\sum_{k=1}^{n-1} H_k^2 = nH_n^2 - (2n+1)H_n + 2n$

(6) $\sum_{k=1}^{n-1} H_k^{(2)} = nH_n^{(2)} - H_n$

(7) $\sum_{k=1}^{n-1} H_kH_{n-k} = (n+1)(H_{n+1}^2 - H_{n+1}^{(2)} - 2H_{n+1} + 2)$

(8) $\sum_{k=1}^{n-1} kH_kH_{n-k} = \binom{n+1}{2}(H_{n+1}^2 - H_{n+1}^{(2)} - 2H_{n+1} + 2)$

Identities (1)–(6) are well known and easily proved by induction on n: see, for instance, the chapters on harmonic numbers in Knuth's classical textbooks [6, §6.3, 6.4] and [9, §1.2.7]. Identities (7) and (8) are proved in [20, Thms. 1,2].

3 Recursive shape indices

A *recursive shape index for phylogenetic trees* [10] is a mapping I that associates to each phylogenetic tree T a real number $I(T) \in \mathbb{R}$ satisfying the following two conditions:

(a) It is invariant under tree isomorphisms and relabelings of leaves.

(b) There exists a symmetrical mapping $f_I : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ such that, for every phylogenetic trees T, T' on disjoint sets of taxa S, S', respectively,

$$I(T\,\widehat{\ }\,T') = I(T) + I(T') + f_I(|S|, |S'|).$$

As we shall see in later sections, the balance indices considered in this paper are recursive shape indices. The following two results extract a common part of the computation of their variances. In them, and henceforth, E_Y applied to a random variable will mean the expected value of this random variable under the Yule model.

Lemma 3. *Let I be a recursive shape index for phylogenetic trees. For every $n \geqslant 1$, let I_n be the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes $I(T)$. Then,*

$$E_Y(I_n^2) = \frac{1}{n-1} \sum_{k=1}^{n-1} \Big(2E_Y(I_k^2) + 4f_I(k, n-k)E_Y(I_k) + 2E_Y(I_k)E_Y(I_{n-k}) + f_I(k, n-k)^2 \Big).$$

Proof. We compute $E_Y(I_n^2)$ using its very definition and Lemma 1. Recall that every tree in $\mathcal{T}_n$ is obtained *twice* as $T_k\,\widehat{\ }\,T'_{n-k}$, for some $1 \leqslant k \leqslant n-1$, some subset $S_k \subseteq \{1, \ldots, n\}$ with k elements, some tree T_k on S_k and some tree T'_{n-k} on S_k^c.

$$\begin{aligned}
E_Y(I_n^2) &= \sum_{T \in \mathcal{T}_n} I(T)^2 \cdot p_Y(T) \\
&= \frac{1}{2} \sum_{k=1}^{n-1} \sum_{\substack{S_k \subsetneq \{1,\ldots,n\} \\ |S_k| = k}} \sum_{T_k \in \mathcal{T}(S_k)} \sum_{T'_{n-k} \in \mathcal{T}(S_k^c)} I(T_k\,\widehat{\ }\,T'_{n-k})^2 \cdot p_Y(T_k\,\widehat{\ }\,T'_{n-k}) \\
&= \frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} \sum_{T_k \in \mathcal{T}_k} \sum_{T'_{n-k} \in \mathcal{T}_{n-k}} \big(I(T_k) + I(T'_{n-k}) + f_I(k, n-k) \big)^2 \frac{2}{(n-1)\binom{n}{k}} P_Y(T_k) P_Y(T'_{n-k}) \\
&= \frac{1}{n-1} \sum_{k=1}^{n-1} \sum_{T_k} \sum_{T'_{n-k}} \big(I(T_k)^2 + I(T'_{n-k})^2 + f_I(k, n-k)^2 + 2I(T_k)I(T'_{n-k}) \\
&\qquad + 2f_I(k, n-k)I(T_k) + 2f_I(k, n-k)I(T'_{n-k}) \big) P_Y(T_k) P_Y(T'_{n-k})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n-1}\sum_{k=1}^{n-1}\Big(\sum_{T_k}\sum_{T'_{n-k}} I(T_k)^2 P_Y(T_k)P_Y(T'_{n-k})\\
&\quad +\sum_{T_k}\sum_{T'_{n-k}} I(T'_{n-k})^2 P_Y(T_k)P_Y(T'_{n-k})\\
&\quad +\sum_{T_k}\sum_{T'_{n-k}} f_I(k,n-k)^2 P_Y(T_k)P_Y(T'_{n-k})\\
&\quad +2\sum_{T_k}\sum_{T'_{n-k}} f_I(k,n-k)I(T_k) P_Y(T_k)P_Y(T'_{n-k})\\
&\quad +2\sum_{T_k}\sum_{T'_{n-k}} f_I(k,n-k)I(T'_{n-k}) P_Y(T_k)P_Y(T'_{n-k})\\
&\quad +2\sum_{T_k}\sum_{T'_{n-k}} I(T_k)I(T'_{n-k}) P_Y(T_k)P_Y(T'_{n-k})\Big)\\
&= \frac{1}{n-1}\sum_{k=1}^{n-1}\Big(\sum_{T_k} I(T_k)^2 P_Y(T_k)+\sum_{T'_{n-k}} I(T'_{n-k})^2 P_Y(T'_{n-k})+f_I(k,n-k)^2\\
&\quad +2\sum_{T_k} f_I(k,n-k)I(T_k)P_Y(T_k)+2\sum_{T'_{n-k}} f_I(k,n-k)I(T'_{n-k})P_Y(T'_{n-k})\\
&\quad +2\Big(\sum_{T_k} I(T_k)P_Y(T_k)\Big)\Big(\sum_{T'_{n-k}} I(T'_{n-k})P_Y(T'_{n-k})\Big)\Big)\\
&= \frac{1}{n-1}\sum_{k=1}^{n-1}\Big(E_Y(I_k^2)+E_Y(I_{n-k}^2)+f_I(k,n-k)^2\\
&\quad +2f_I(k,n-k)(E_Y(I_k)+E_Y(I_{n-k}))+2E_Y(I_k)\cdot E_Y(I_{n-k})\Big)\\
&= \frac{1}{n-1}\sum_{k=1}^{n-1}(2E_Y(I_k^2)+4f_I(k,n-k)E_Y(I_k)+2E_Y(I_k)\cdot E_Y(I_{n-k})+f_I(k,n-k)^2)
\end{aligned}
$$

as we claimed. □

Corollary 1. *Let I be a recursive shape index for phylogenetic trees and, for every $n \geqslant 1$, let I_n be the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes $I(T)$. Set*

$$
\begin{aligned}
&\varepsilon_I(a,b-1) = f_I(a,b)-f_I(a,b-1) \text{ for every } a\geqslant 1 \text{ and } b\geqslant 2\\
&R_I(n-1) = E_Y(I_n)-E_Y(I_{n-1}) \text{ for every } n\geqslant 2
\end{aligned}
$$

If $E_Y(I_1)=0$, then

$$
\begin{aligned}
E_Y(I_n^2) &= \frac{n}{n-1}E_Y(I_{n-1}^2)+\frac{4}{n-1}\sum_{k=1}^{n-2}\varepsilon(k,n-1-k)\cdot E_Y(I_k)\\
&\quad +\frac{4}{n-1}f_I(n-1,1)E_Y(I_{n-1})+\frac{2}{n-1}\sum_{k=1}^{n-2}E_Y(I_k)R(n-k-1)\\
&\quad +\frac{f_I(n-1,1)^2}{n-1}+\frac{1}{n-1}\sum_{k=1}^{n-2}(f_I(k,n-k)^2-f_I(k,n-k-1)^2).
\end{aligned}
$$

Proof. By Lemma 3,

$$\begin{aligned}E_Y(I_n^2) = {} & \frac{2}{n-1}\sum_{k=1}^{n-1} E_Y(I_k^2) + \frac{4}{n-1}\sum_{k=1}^{n-1} f_I(k,n-k)E_Y(I_k) \\ & + \frac{2}{n-1}\sum_{k=1}^{n-1} E_Y(I_k)E_Y(I_{n-k}) + \frac{1}{n-1}\sum_{k=1}^{n-1} f_I(k,n-k)^2,\end{aligned}$$

and in particular

$$\begin{aligned}E_Y(I_{n-1}^2) = {} & \frac{2}{n-2}\sum_{k=1}^{n-2} E_Y(I_k^2) + \frac{4}{n-2}\sum_{k=1}^{n-2} f_I(k,n-1-k)E_Y(I_k) \\ & + \frac{2}{n-2}\sum_{k=1}^{n-2} E_Y(I_k)E_Y(I_{n-1-k}) + \frac{1}{n-2}\sum_{k=1}^{n-2} f_I(k,n-k-1)^2.\end{aligned}$$

Therefore

$$\begin{aligned}E_Y(I_n^2) = {} & \frac{n-2}{n-1}\cdot\frac{2}{n-2}\sum_{k=1}^{n-2} E_Y(I_k^2) + \frac{2}{n-1}E_Y(I_{n-1}^2) \\ & + \frac{n-2}{n-1}\cdot\frac{4}{n-2}\sum_{k=1}^{n-2} E_Y(I_k)(f_I(k,n-1-k) + \varepsilon_I(k,n-1-k)) \\ & \quad + \frac{4}{n-1} f_I(n-1,1)E_Y(I_{n-1}) \\ & + \frac{n-2}{n-1}\cdot\frac{2}{n-2}\sum_{k=1}^{n-2} E_Y(I_k)(E_Y(I_{n-1-k}) + R_I(n-k-1)) \\ & \quad + \frac{2}{n-1}E_Y(I_{n-1})E_Y(I_1) \\ & + \frac{n-2}{n-1}\cdot\frac{1}{n-2}\sum_{k=1}^{n-2} f_I(k,n-k-1)^2 + \frac{1}{n-1}\sum_{k=1}^{n-1} f_I(k,n-k)^2 \\ & \quad - \frac{n-2}{n-1}\cdot\frac{1}{n-2}\sum_{k=1}^{n-2} f_I(k,n-k-1)^2 \\ = {} & \frac{n-2}{n-1}E_Y(I_{n-1}^2) + \frac{2}{n-1}E_Y(I_{n-1}^2) + \frac{4}{n-1}\sum_{k=1}^{n-2}\varepsilon_I(k,n-1-k)\cdot E_Y(I_k) \\ & + \frac{4}{n-1} f_I(n-1,1)E_Y(I_{n-1}) + \frac{2}{n-1}\sum_{k=1}^{n-2} E_Y(I_k)R_I(n-k-1) \\ & + \frac{1}{n-1}\sum_{k=1}^{n-1} f_I(k,n-k)^2 - \frac{1}{n-1}\sum_{k=1}^{n-2} f_I(k,n-k-1)^2 \\ = {} & \frac{n}{n-1}E_Y(I_{n-1}^2) + \frac{4}{n-1}\sum_{k=1}^{n-2}\varepsilon_I(k,n-1-k)\cdot E_Y(I_k) \\ & + \frac{4}{n-1} f_I(n-1,1)E_Y(I_{n-1}) + \frac{2}{n-1}\sum_{k=1}^{n-2} E_Y(I_k)R_I(n-k-1) \\ & + \frac{1}{n-1}\sum_{k=1}^{n-2}(f_I(k,n-k)^2 - f_I(k,n-k-1)^2) + \frac{1}{n-1}\cdot f_I(n-1,1)^2.\end{aligned}$$

□

4 The variance of Sackin's index

The *Sackin index* of a phylogenetic tree $T \in \mathcal{T}_n$ is defined as the sum of the depths of its leaves:

$$S(T) = \sum_{i=1}^{n} \delta_T(i).$$

It is well known (cf. [13, Eq. (6)]) that if $T_k \in \mathcal{T}(S_k)$, for some $\emptyset \neq S_k \subsetneq \{1,\dots,n\}$, and $T'_{n-k} \in \mathcal{T}(S_k^c)$, then

$$S(T_k \widehat{\ } T'_{n-k}) = S(T_k) + S(T'_{n-k}) + n.$$

Let S_n be the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes $S(T)$. Its expected value under the Yule model is [8]

$$E_Y(S_n) = 2n(H_n - 1).$$

In particular $E_Y(S_1) = 0$. Notice that $E_Y(S_n)$ satisfies the recurrence

$$E_Y(S_{n+1}) = E_Y(S_n) + 2H_n.$$

Indeed,

$$\begin{aligned} E_Y(S_{n+1}) - E_Y(S_n) &= 2(n+1)(H_{n+1} - 1) - 2n(H_n - 1) \\ &= 2(n+1)(H_n + \frac{1}{n+1} - 1) - 2n(H_n - 1) = 2H_n. \end{aligned}$$

In this section we compute the variance of S_n under this model.

Theorem 1. $E_Y(S_n^2) = 4n^2(H_n^2 - H_n^{(2)} - 2H_n) - 2nH_n + 11n^2 - n$

Proof. As we have seen, Sackin's index satisfies the hypotheses in Corollary 1, with $f_S(k, n-k) = n$, and hence $\varepsilon_S(k, n-k-1) = 1$, and $R_S(k) = 2H_k$. Therefore

$$\begin{aligned} E_Y(S_n^2) = {} & \frac{n}{n-1} E_Y(S_{n-1}^2) + \frac{4}{n-1} \sum_{k=1}^{n-2} E_Y(S_k) \\ & + \frac{4}{n-1} n E_Y(S_{n-1}) + \frac{2}{n-1} \sum_{k=1}^{n-2} E_Y(S_k) 2H_{n-k-1} \\ & + \frac{n^2}{n-1} + \frac{1}{n-1} \sum_{k=1}^{n-2} (n^2 - (n-1)^2) \\ = {} & \frac{n}{n-1} E_Y(S_{n-1}^2) + \frac{4}{n-1} \sum_{k=1}^{n-2} E_Y(S_k) + 8n(H_{n-1} - 1) \\ & + \frac{4}{n-1} \sum_{k=1}^{n-2} E_Y(S_k) H_{n-k-1} + 3n - 2 \end{aligned}$$

Now, by Lemma 2,

$$\begin{aligned} \frac{4}{n-1} \sum_{k=1}^{n-2} E_Y(S_k) &= \frac{8}{n-1} \sum_{k=1}^{n-2} k(H_k - 1) \\ &= \frac{8}{n-1} \Big(\frac{1}{4}(n-1)(n-2)(2H_{n-1} - 1) - \frac{1}{2}(n-1)(n-2) \Big) \\ &= 2(n-2)(2H_{n-1} - 3) \end{aligned}$$

$$\begin{aligned}
\frac{4}{n-1}\sum_{k=1}^{n-2} E_Y(S_k)H_{n-k-1} &= \frac{8}{n-1}\sum_{k=1}^{n-2} k(H_k-1)H_{n-k-1}\\
&= \frac{8}{n-1}\sum_{k=1}^{n-2} kH_kH_{n-k-1} - \frac{8}{n-1}\sum_{k=1}^{n-2} kH_{n-k-1}\\
&= \frac{8}{n-1}\sum_{k=1}^{n-2} kH_kH_{n-k-1} - \frac{8}{n-1}\sum_{k=1}^{n-2} (n-k-1)H_k\\
&= 4n(H_n^2-H_n^{(2)}-2H_n+2) - 8\sum_{k=1}^{n-2} H_k + \frac{8}{n-1}\sum_{k=1}^{n-2} kH_k\\
&= 4n(H_n^2-H_n^{(2)}-2H_n+2) - 8(n-1)(H_{n-1}-1) + 2(n-2)(2H_{n-1}-1)\\
&= 4n(H_n^2-H_n^{(2)}-2H_{n-1}-2\cdot\frac{1}{n}+2) - 4nH_{n-1} + 6n-4\\
&= 4n(H_n^2-H_n^{(2)}-3H_{n-1}) + 14n-12
\end{aligned}$$

And thus

$$\begin{aligned}
E_Y(S_n^2) &= \frac{n}{n-1}E_Y(S_{n-1}^2) + 2(n-2)(2H_{n-1}-3) + 8n(H_{n-1}-1)\\
&\qquad +4n(H_n^2-H_n^{(2)}-3H_{n-1}) + 14n-12+3n-2\\
&= \frac{n}{n-1}E_Y(S_{n-1}^2) + 4n(H_n^2-H_n^{(2)}) - 8H_{n-1} + 3n-2
\end{aligned}$$

Setting $x_n = E_Y(S_n^2)/n$, this equation becomes

$$x_n = x_{n-1} + 4(H_n^2-H_n^{(2)}) - 8\frac{H_{n-1}}{n} + 3 - \frac{2}{n}$$

The solution of this equation with $x_1 = 0$ is

$$\begin{aligned}
x_n &= \sum_{k=2}^{n}\Big(4(H_k^2-H_k^{(2)}) - 8\frac{H_{k-1}}{k} + 3 - \frac{2}{k}\Big)\\
&= 4\sum_{k=2}^{n}(H_k^2-H_k^{(2)}) - 8\sum_{k=1}^{n-1}\frac{H_k}{k+1} + 3(n-1) - 2\sum_{k=2}^{n}\frac{1}{k}\\
&= 4\sum_{k=2}^{n}(H_k^2-H_k^{(2)}) - 4(H_n^2-H_n^{(2)}) + 3(n-1) - 2(H_n-1)\\
&= 4\sum_{k=2}^{n-1}(H_k^2-H_k^{(2)}) - 2H_n + 3n-1\\
&= 4(nH_n^2-(2n+1)H_n+2n-nH_n^{(2)}+H_n) - 2H_n + 3n-1\\
&= 4n(H_n^2-H_n^{(2)}-2H_n) - 2H_n + 11n-1
\end{aligned}$$

and therefore

$$E_Y(S_n^2) = nx_n = 4n^2(H_n^2-H_n^{(2)}-2H_n) - 2nH_n + 11n^2 - n$$

as we claimed. □

Corollary 2. *The variance of S_n under the Yule model is*

$$\sigma_Y^2(S_n) = 7n^2 - 4n^2H_n^{(2)} - 2nH_n - n.$$

Proof. It is obtained by replacing in the formula $\sigma_Y^2(S_n) = E_Y(S_n^2) - E_Y(S_n)^2$ the value of $E_Y(S_n^2)$ obtained in the last theorem and $E_Y(S_n) = 2n(H_n - 1)$. □

From this exact formula we can obtain an $O(1/n)$ approximation of $\sigma_Y^2(S_n)$, which refines the limit formula obtained in [1].

Corollary 3. $\sigma_Y^2(S_n) = \Big(7 - \dfrac{2\pi^2}{3}\Big)n^2 + n(3 - 2\ln(n) - 2\gamma) - 3 + O\Big(\dfrac{1}{n}\Big)$.

5 The variance of the total cophenetic index Φ

The *total cophenetic index* of a phylogenetic tree $T \in \mathcal{T}_n$ is defined as the sum of the cophenetic values of its pairs of leaves:

$$\Phi(T) = \sum_{1 \leqslant i < j \leqslant n} \varphi_T(i,j).$$

By [11, Lem. 4], if $T_k \in \mathcal{T}(S_k)$, for some $\emptyset \neq S_k \subsetneq \{1,\dots,n\}$ with k elements, and $T'_{n-k} \in \mathcal{T}(S_k^c)$, then

$$\Phi(T_k\widehat{\ }T_{n-k}) = \Phi(T_k) + \Phi(T_{n-k}) + \binom{k}{2} + \binom{n-k}{2}.$$

Therefore, Φ is a recursive shape index with $f_\Phi(k, n-k) = \binom{k}{2} + \binom{n-k}{2}$, and in particular $\varepsilon_\Phi(k, n-k-1) = n-k-1$.

Let Φ_n be the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes its total cophenetic index $\Phi(T)$. The expected value under the Yule model of Φ_n is [11]

$$E_Y(\Phi_n) = n(n-1) - 2n(H_n - 1) = n(n + 1 - 2H_n).$$

In particular, $E_Y(\Phi_1) = 0$. This expected value satisfies the recurrence

$$E_Y(\Phi_n) = E_Y(\Phi_{n-1}) + 2(n - 1 - H_{n-1}),$$

and therefore $R(k) = 2(k - H_k)$.

In this section we compute the variance of Φ_n under the Yule model.

Theorem 2. $E_Y(\Phi_n^2) = 4n^2(H_n^2 - H_n^{(2)}) - 2(2n^3 + 2n^2 + 3n)H_n + \frac{1}{12}(13n^4 + 14n^3 + 143n^2 - 2n)$

Proof. Φ satisfies the hypothesis of Corollary 1, with

$$\varepsilon_\Phi(k, n-k-1) = n-k-1, \quad R(k) = 2(k - H_k).$$

Therefore, by the aforementioned result,

$$\begin{aligned}
E_Y(\Phi_n^2) = {} & \frac{n}{n-1}E_Y(\Phi_{n-1}^2) + \frac{4}{n-1}\sum_{k=1}^{n-2}(n-k-1)E_Y(\Phi_k) + \frac{4}{n-1}\binom{n-1}{2}E_Y(\Phi_{n-1}) \\
& + \frac{2}{n-1}\sum_{k=1}^{n-2} 2E_Y(\Phi_k)((n-k-1) - H_{n-k-1}) + \frac{1}{n-1}\binom{n-1}{2}^2 \\
& + \frac{1}{n-1}\sum_{k=1}^{n-2}\left(\left(\binom{k}{2} + \binom{n-k}{2}\right)^2 - \left(\binom{k}{2} + \binom{n-k-1}{2}\right)^2\right) \\
= {} & \frac{n}{n-1}E_Y(\Phi_{n-1}^2) + \frac{8}{n-1}\sum_{k=1}^{n-2}(n-k-1)(k^2+k-2kH_k) + 2(n-2)(n-1)(n-2H_{n-1}) \\
& - \frac{4}{n-1}\sum_{k=1}^{n-2} H_{n-k-1}(k^2+k-2kH_k) + \frac{1}{12}(n-2)(7n^2-21n+12) \\
= {} & \frac{n}{n-1}E_Y(\Phi_{n-1}^2) - 4(n-2)(n-1)H_{n-1} \\
& - 16\sum_{k=1}^{n-2} kH_k + \frac{16}{n-1}\sum_{k=1}^{n-2} k^2H_k \\
& - \frac{4}{n-1}\sum_{k=1}^{n-2} k^2H_{n-k-1} - \frac{4}{n-1}\sum_{k=1}^{n-2} kH_{n-k-1} + \frac{8}{n-1}\sum_{k=1}^{n-2} kH_kH_{n-k-1} \\
& + \frac{1}{12}(n-2)(39n^2-37n+12) \\
= {} & \frac{n}{n-1}E_Y(\Phi_{n-1}^2) - 4(n-2)(n-1)H_{n-1} \\
& - 16\sum_{k=1}^{n-2} kH_k + \frac{16}{n-1}\sum_{k=1}^{n-2} k^2H_k \\
& - \frac{4}{n-1}\sum_{k=1}^{n-2}(n-k-1)^2H_k - \frac{4}{n-1}\sum_{k=1}^{n-2}(n-k-1)H_k + \frac{8}{n-1}\sum_{k=1}^{n-2} kH_kH_{n-k-1} \\
& + \frac{1}{12}(n-2)(39n^2-37n+12) \\
= {} & \frac{n}{n-1}E_Y(\Phi_{n-1}^2) - 4(n-2)(n-1)H_{n-1} \\
& - 16\sum_{k=1}^{n-2} kH_k + \frac{16}{n-1}\sum_{k=1}^{n-2} k^2H_k + \frac{8}{n-1}\sum_{k=1}^{n-2} kH_kH_{n-k-1} \\
& - 4(n-1)\sum_{k=1}^{n-2} H_k + 8\sum_{k=1}^{n-2} kH_k - \frac{4}{n-1}\sum_{k=1}^{n-2} k^2H_k - 4\sum_{k=1}^{n-2} H_k + \frac{4}{n-1}\sum_{k=1}^{n-2} kH_k \\
& + \frac{1}{12}(n-2)(39n^2-37n+12) \\
= {} & \frac{n}{n-1}E_Y(\Phi_{n-1}^2) - 4(n-2)(n-1)H_{n-1} \\
& + \frac{12}{n-1}\sum_{k=1}^{n-2} k^2H_k + \frac{12-8n}{n-1}\sum_{k=1}^{n-2} kH_k - 4n\sum_{k=1}^{n-2} H_k + \frac{8}{n-1}\sum_{k=1}^{n-2} kH_kH_{n-k-1} \\
& + \frac{1}{12}(n-2)(39n^2-37n+12)
\end{aligned}$$

$$\begin{aligned}
&= \frac{n}{n-1}E_Y(\Phi^2_{n-1}) - 4(n-2)(n-1)H_{n-1}\\
&\quad + \frac{12}{n-1}\cdot\frac{1}{36}(n-1)(n-2)\Big((12n-18)H_{n-1}-4n+3\Big)\\
&\quad + \frac{12-8n}{n-1}\cdot\frac{1}{4}(n-1)(n-2)(2H_{n-1}-1) - 4n(n-1)(H_{n-1}-1)\\
&\quad + \frac{8}{n-1}\binom{n}{2}(H_n^2 - H_n^{(2)} - 2H_n + 2) + \frac{1}{12}(n-2)(39n^2-37n+12)\\
&= \frac{n}{n-1}E_Y(\Phi^2_{n-1}) + 4n(H_n^2 - H_n^{(2)}) - 8nH_n\\
&\quad -8(n-1)^2H_{n-1} + \frac{1}{12}(39n^3-59n^2+94n+24)\\
&= \frac{n}{n-1}E_Y(\Phi^2_{n-1}) + 4n(H_n^2 - H_n^{(2)}) - 8(n^2-n+1)H_{n-1} + \frac{1}{12}(39n^3-59n^2+94n-72)
\end{aligned}$$

Setting $x_n = E_Y(\Phi^2_n)/n$, this equation becomes

$$x_n = x_{n-1} + 4(H_n^2 - H_n^{(2)}) - 8\Big(n-1+\frac{1}{n}\Big)H_{n-1} + \frac{1}{12}\Big(39n^2-59n+94-\frac{72}{n}\Big)$$

The solution of this equation with $x_1 = 0$ is

$$\begin{aligned}
x_n &= \sum_{k=2}^{n}\Big(4(H_k^2 - H_k^{(2)}) - 8\Big(k-1+\frac{1}{k}\Big)H_{k-1} + \frac{1}{12}\Big(39k^2-59k+94-\frac{72}{k}\Big)\Big)\\
&= 4\sum_{k=2}^{n}H_k^2 - 4\sum_{k=2}^{n}H_k^{(2)} - 8\sum_{k=1}^{n-1}kH_k - 8\sum_{k=1}^{n-1}\frac{H_k}{k+1} + \frac{1}{12}\sum_{k=2}^{n}\Big(39k^2-59k+94-\frac{72}{k}\Big)\\
&= 4\sum_{k=2}^{n-1}H_k^2 - 4\sum_{k=2}^{n-1}H_k^{(2)} - 8\sum_{k=1}^{n-1}kH_k + \frac{1}{12}\sum_{k=2}^{n}\Big(39k^2-59k+94-\frac{72}{k}\Big)\\
&= 4n(H_n^2 - H_n^{(2)}) - 8nH_n + 8n - 2n(n-1)(2H_n-1) - 6H_n + \frac{1}{12}(13n^3-10n^2+71n-2)\\
&= 4n(H_n^2 - H_n^{(2)}) - 2(2n^2+2n+3)H_n + \frac{1}{12}(13n^3+14n^2+143n-2)
\end{aligned}$$

Therefore

$$E_Y(\Phi^2_n) = nx_n = 4n^2(H_n^2 - H_n^{(2)}) - 2(2n^3+2n^2+3n)H_n + \frac{1}{12}(13n^4+14n^3+143n^2-2n)$$

as we claimed. □

Corollary 4. $\sigma^2_Y(\Phi_n) = \frac{1}{12}(n^4-10n^3+131n^2-2n) - 4n^2H_n^{(2)} - 6nH_n$

Proof. Simply replace in the formula $\sigma^2_Y(\Phi_n) = E_Y(\Phi^2_n) - E_Y(\Phi_n)^2$ the value of $E_Y(\Phi^2_n)$ obtained in the last theorem and the value of $E_Y(\Phi_n)$ recalled above. □

Corollary 5. $\sigma^2_Y(\Phi_n) = \dfrac{1}{12}n^4 - \dfrac{5}{6}n^3 + \Big(\dfrac{131}{12} - \dfrac{2\pi^2}{3}\Big)n^2 - 6n\ln(n) + \Big(\dfrac{23}{6} - \gamma\Big)n - 5 + O\Big(\dfrac{1}{n}\Big).$

6 The variance of $\overline{\Phi}$

For every $T \in \mathcal{T}_n$, let

$$\overline{\Phi}(T) = S(T) + \Phi(T) = \sum_{1\leqslant i\leqslant j\leqslant n}\varphi_T(i,j)$$

Lemma 4. *If $T_k \in \mathcal{T}(S_k)$, with $\emptyset \neq S_k \subsetneq \{1,\dots,n\}$ and $|S_k| = k$, and $T'_{n-k} \in \mathcal{T}(S_k^c)$, then*

$$\overline{\Phi}(T_k\,\hat{}\,T_{n-k}) = \overline{\Phi}(T_k) + \overline{\Phi}(T_{n-k}) + \binom{k+1}{2} + \binom{n-k+1}{2}.$$

Proof. Since $S(T_k\,\hat{}\,T_{n-k}) = S(T_k) + S(T_{n-k}) + n$ and

$$\Phi(T_k\,\hat{}\,T_{n-k}) = \Phi(T_k) + \Phi(T_{n-k}) + \binom{k}{2} + \binom{n-k}{2},$$

we have that

$$\begin{aligned}\overline{\Phi}(T_k\,\hat{}\,T_{n-k}) &= \overline{\Phi}(T_k) + \overline{\Phi}(T_{n-k}) + \binom{k}{2} + \binom{n-k}{2} + n\\ &= \overline{\Phi}(T_k) + \overline{\Phi}(T_{n-k}) + \binom{k+1}{2} + \binom{n-k+1}{2}.\end{aligned}$$

□

So, $\overline{\Phi}$ is a recursive shape index for phylogenetic trees with $f_{\overline{\Phi}}(a,b) = \binom{a+1}{2} + \binom{b+1}{2}$, and hence $\varepsilon_{\overline{\Phi}}(a,b) = b+1$.

Let $\overline{\Phi}_n$ be the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes $\overline{\Phi}(T)$. Its expected value under the Yule model is [11]

$$E_Y(\overline{\Phi}_n) = n(n-1).$$

In particular, $E_Y(\overline{\Phi}_1) = 0$ and $R_{\overline{\Phi}}(k) = 2k$. In this section we compute the variance of $\overline{\Phi}_n$.

Theorem 3. $E_Y(\overline{\Phi}_n^2) = \frac{1}{12}(13n^4 - 30n^3 + 23n^2 - 6n)$.

Proof. $\overline{\Phi}$ is a recursive shape index for phylogenetic trees with

$$\varepsilon_{\overline{\Phi}}(k, n-k-1) = n-k, \quad R(k) = 2(k).$$

Then, by Corollary 1,

$$\begin{aligned}E_Y(\overline{\Phi}_n^2) &= \frac{n}{n-1}E_Y(\overline{\Phi}_{n-1}^2) + \frac{4}{n-1}\sum_{k=1}^{n-2}(n-k)E_Y(\overline{\Phi}_k) + \frac{4}{n-1}\Big(\binom{n}{2}+1\Big)E_Y(\overline{\Phi}_{n-1})\\ &\quad+\frac{2}{n-1}\sum_{k=1}^{n-2}2E_Y(S_k)(n-k-1) + \frac{1}{n-1}\Big(\binom{n}{2}+\binom{2}{2}\Big)^2\\ &\quad+\frac{1}{n-1}\sum_{k=1}^{n-2}\Big(\Big(\binom{k+1}{2}+\binom{n-k+1}{2}\Big)^2 - \Big(\binom{k+1}{2}+\binom{n-k}{2}\Big)^2\Big)\\ &= \frac{n}{n-1}E_Y(\overline{\Phi}_{n-1}^2)\\ &\quad+\frac{4}{n-1}\sum_{k=1}^{n-2}(n-k)k(k-1) + \frac{4}{n-1}\Big(\binom{n}{2}+1\Big)(n-1)(n-2)\\ &\quad+\frac{4}{n-1}\sum_{k=1}^{n-2}k(k-1)(n-k-1)\\ &\quad+\frac{1}{n-1}\sum_{k=1}^{n-2}(2(n-k)\Big(\binom{k+1}{2}+\binom{n-k}{2}\Big) + (n-k)^2) + \frac{1}{n-1}\Big(\binom{n}{2}+1\Big)^2\\ &= \frac{n}{n-1}E_Y(\overline{\Phi}_{n-1}^2) + \frac{1}{4}n(13n^2 - 33n + 22)\end{aligned}$$

Setting $x_n = E_Y(\overline{\Phi}_n^2)/n$, this recurrence becomes

$$x_n = x_{n-1} + \frac{1}{4}(13n^2 - 33n + 22)$$

and the solution with $x_1 = 0$ is

$$x_n = \frac{1}{12}(13n^3 - 30n^2 + 23n - 6)$$

from where we deduce that

$$E_Y(\overline{\Phi}_n^2) = nx_n = \frac{1}{12}(13n^4 - 30n^3 + 23n^2 - 6n)$$

as we claimed. □

Corollary 6. *The variance of $\overline{\Phi}_n$ under the Yule model is*

$$\sigma_Y^2(\overline{\Phi}_n) = 2\binom{n}{4}$$

Proof. Simply apply that $\sigma_Y^2(\overline{\Phi}_n) = E_Y(\overline{\Phi}_n^2) - E_Y(\overline{\Phi}_n)^2$. □

7 The covariances

In this section we obtain the covariance under the Yule model of S_n and Φ_n from the formulas obtained in the previous sections for $E_Y(\overline{\Phi}_n^2)$, $E_Y(S_n^2)$ and $E_Y(\Phi_n^2)$.

Corollary 7. $cov_Y(S_n, \Phi_n) = 4n(nH_n^{(2)} + H_n) + \frac{1}{6}n(n^2 - 51n + 2)$.

Proof. Recall that $cov_Y(S_n, \Phi_n) = E_Y(S_n \cdot \Phi_n) - E_Y(S_n) \cdot E_Y(\Phi_n)$. Now

$$E_Y(\overline{\Phi}_n^2) = E_Y((S_n + \Phi_n)^2) = E_Y(S_n^2) + E_Y(\Phi_n^2) + 2E_Y(S_n \cdot \Phi_n)$$

and therefore

$$E_Y(S_n \cdot \Phi_n) = \frac{1}{2}(E_Y(\overline{\Phi}_n^2) - E_Y(S_n^2) - E_Y(\Phi_n^2))$$

from where we obtain, replacing $E_Y(\overline{\Phi}_n^2)$, $E_Y(S_n^2)$ and $E_Y(\Phi_n^2)$ by their values obtained in the previous sections, that

$$E_Y(S_n \cdot \Phi_n) = 2n(n^2 + 3n + 2)H_n - 4n^2(H_n^2 - H_n^{(2)}) - \frac{11}{6}n^3 - \frac{21}{2}n^2 + \frac{1}{3}n.$$

Subtracting $E_Y(S_n) \cdot E_Y(\Phi_n) = 2n^2(H_n - 1)(n + 1 - 2H_n)$ to this expression, we finally obtain the formula in the statement. □

Corollary 8. $cov_Y(S_n, \overline{\Phi}_n) = 2nH_n + \dfrac{1}{6}n(n^2 - 9n - 4)$

Proof. By the bilinearity of covariances, $cov_Y(S_n, \overline{\Phi}_n) = cov_Y(S_n, S_n + \Phi_n) = \sigma_Y^2(S_n) + cov_Y(S_n, \Phi_n)$. □

Corollary 9.

$$\begin{aligned} cov_Y(S_n,\mathit{\Phi}_n) &= \frac{1}{6}n^3+\Big(\frac{2\pi^2}{3}-\frac{17}{2}\Big)n^2+4n\ln(n)+\frac{1}{3}(12\gamma-11)n+4+O\Big(\frac{1}{n}\Big)\\ cov_Y(S_n,\overline{\mathit{\Phi}}_n) &= \frac{1}{6}n^3-\frac{3}{2}n^2+2n\ln(n)+\frac{1}{3}(6\gamma-2)n+1+O\Big(\frac{1}{n}\Big) \end{aligned}$$

From the formulas for $\sigma_Y^2(\mathit{\Phi}_n)$, $\sigma_Y^2(\overline{\mathit{\Phi}}_n)$, and $cov_Y(S_n,\mathit{\Phi}_n)$, we can compute Pearson's correlation coefficient between S_n and $\mathit{\Phi}_n$,

$$cor_Y(S_n,\mathit{\Phi}_n)=\frac{cov_Y(S_n,\mathit{\Phi}_n)}{\sqrt{\sigma_Y^2(\mathit{\Phi}_n)\cdot\sigma_Y^2(\overline{\mathit{\Phi}}_n)}}.$$

The exact formula for this coefficient is

$$\begin{aligned} &cor_Y(S_n,\mathit{\Phi}_n)\\ &=\frac{4n(nH_n^{(2)}+H_n)+\frac{1}{6}n(n^2-51n+2)}{\sqrt{\Big(7n^2-4n^2H_n^{(2)}-2nH_n-n\Big)\Big(\frac{1}{12}(n^4-10n^3+131n^2-2n)-4n^2H_n^{(2)}-6nH_n\Big)}} \end{aligned}$$

and in the limit it is equal to

$$cor_Y(S_n,\mathit{\Phi}_n)\sim\frac{1}{6\sqrt{\Big((7-\frac{2\pi^2}{3})\cdot\frac{1}{12}\Big)}}=0.89059$$

8 Conclusions

In this paper we have obtained exact formulas for the variance under the Yule model of the Sackin index S, the total cophenetic index $\mathit{\Phi}$ and their sum $\overline{\mathit{\Phi}}$, and for the covariances of S and $\mathit{\Phi},\overline{\mathit{\Phi}}$. Unlike other expressions published so far in the literature, our formulas are valid on spaces $\mathcal{T}_n$ of binary phylogenetic trees with any number n of leaves, and not only asymptotic formulas for large such n, and they are not recursive, but explicit.

The proofs consist of elementary, although long and involved, algebraic computations. Since it is not difficult to slip some mistake in such long algebraic computations, to double-check the results we have directly computed these variances and covariances on $\mathcal{T}_n$ for $n=3,\dots,9$ and confirmed that our formulas give the right results. The values obtained are given in the next table. The Python scripts used to compute them are available from the authors.

	3	4	5	6	7	8	9
$\sigma_Y^2(S_n)$	0	0.222222	0.805556	1.84	3.877778	5.49424	8.193827
$\sigma_Y^2(\mathit{\Phi}_n)$	0	0.888889	5.138889	17.04	42.787778	90.522812	170.350969
$\sigma_Y^2(\overline{\mathit{\Phi}}_n)$	0	2	10	30	70	140	252
$cov_Y(S_n,\mathit{\Phi}_n)$	0	0.444444	2.0277778	5.56	11.912222	21.991474	36.727602
$cor_Y(S_n,\mathit{\Phi}_n)$	-	1	0.996639	0.992958	0.989408	0.986101	0.983053

Table 1. Values of $\sigma_Y^2(S_n)$, $\sigma_Y^2(\mathit{\Phi}_n)$, $\sigma_Y^2(\overline{\mathit{\Phi}}_n)$, $cov_Y(S_n,\mathit{\Phi}_n)$, and $cor_Y(S_n,\mathit{\Phi}_n)$ for $n=3,\dots,9$. They agree with the values given by our formulas.

It can be seen in this table that the values of the variance of S_n are smaller than those of the variance of $\varPhi$ or $\overline{\varPhi}$. Actually, as we have seen in the text, $\sigma_Y^2(S_n)$ has order $O(n^2)$, while $\sigma_Y^2(\varPhi_n)$ and $\sigma_Y^2(\overline{\varPhi}_n)$ are $O(n^4)$. This is consistent with the fact that $\varPhi$ and $\overline{\varPhi}$ have larger spans of values than S, $O(n^3)$ instead of $O(n^2)$, and much less ties. It is also deduced from the formulas obtained in this paper, and from this table for small values of n, that there is a strong direct linear correlation between S_n and $\varPhi_n$, although in the limit Pearson's coefficient between them decreases to 0.89.

It remains to compute an exact formula for the variance of Colless' index C and of the covariances of C and $S, \varPhi$. The variance can be obtained using again Corollary 1, while the covariances can be obtained using a result similar to that corollary, giving a recurrence for the expected value of the product of two recursive shape indices. In both cases, the computations are much longer and more involved than those included here. We shall report on them elsewhere.

Acknowledgements

The research reported in this paper has been partially supported by the Spanish government and the UE FEDER program, through projects MTM2009-07165 and TIN2008-04487-E/TIN. We thank J. Miró for several comments on a previous version of this work.

References

1. M. G. B. Blum, O. François, S. Janson, The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. Ann. Appl. Probab. 16 (2006), 2195–2214.
2. J. Brown, Probabilities of evolutionary trees. Syst. Biol. 43 (1994), 78–91.
3. L. L. Cavalli-Sforza, A. Edwards, Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet., 19 (1967), 233–257.
4. D. H. Colless, Review of "Phylogenetics: the theory and practice of phylogenetic systematics". Sys. Zool, 31 (1982), 100–104.
5. J. Felsenstein, Inferring Phylogenies. Sinauer Associates Inc., 2004.
6. R. Graham, D. Knuth, O. Patashnik, Concrete Mathematics. Addison-Wesley (1994).
7. E. Harding, The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. 3 (1971), 44–77.
8. S. B. Heard, Patterns in Tree Balance among Cladistic, Phenetic, and Randomly Generated Phylogenetic Trees. Evolution 46 (1992), 1818–1826
9. D. Knuth, The Art of Computer Programming, Vol. 1: Fundamental Algorithms (3rd Edition). Addison-Wesley (1997).
10. F. Matsen, Optimization Over a Class of Tree Shape Statistics. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 4 (2007), 506–512.
11. A. Mir, F. Rosselló, L. Rotger, A new balance index for phylogenetic trees. arXiv:1202.1223v1 [q-bio.PE] (2012), submitted to Math. Biosc.
12. A. Mooers, S. B. Heard, Inferring evolutionary process from phylogenetic tree shape. Quart. Rev. Biol. 72 (1997) 31–54.
13. J. S. Rogers, Central moments and probability distributions of three measures of phylogenetic tree imbalance, Sys. Biol. 45 (1996), 99–110.
14. D. E. Rosen, Vicariant Patterns and Historical Explanation in Biogeography. Syst. Biol. 27 (1978), 159–188.
15. M. J. Sackin, "Good" and "bad" phenograms. Sys. Zool, 21 (1972), 225–226.
16. K.T. Shao, R. Sokal, Tree balance. Sys. Zool, 39 (1990), 226–276.
17. R. Sokal, F. Rohlf, The Comparison of Dendrograms by Objective Methods. Taxon 11 (1962), 33–40.

18. M. Steel, A. McKenzie, Distributions of cherries for two models of trees. Math. Biosc. 164 (2000), 81–92.
19. M. Steel, A. McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models. Math. Biosc. 170 (2001), 91–112.
20. C. Wei, D. Gong, Q. Wang, Chu-Vandermonde convolution and harmonic number identities. arXiv:1201.0420v1 [math.CO] (2012)
21. G. U. Yule, A mathematical theory of evolution based on the conclusions of Dr J. C. Willis. Phil. Trans. Royal Soc. (London) Series B 213 (1924), 21–87.